Briefings in Bioinformatics, 2021, pp. 1–9

doi: DOI
Advance Access Publication Date: Day Month Year
Paper

PAPER

RNALocator: a neural network-based model for prediction of mRNA sub-cellular localization

Negin Sadat Babaiha,[1] Rosa Aghdam[2,*] and Changiz Eslahchi[1,2,*]

[1]Department of Computer and Data Sciences, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran and [2]School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
[*]Corresponding authors. ch-eslahchi@sbu.ac.ir, rosa.aghdam@ipm.ir

Abstract

Localization of RNA is an efficient way to target proteins to a specific region of a cell and is known as a strategy for gene regulation, as well. The importance of correct intracellular RNA placement in the development of embryonic and neural dendrites has also been clearly demonstrated in former studies. Improper localization of RNA in the cell, which has been shown to occur due to a variety of reasons, including mutations in trans-regulatory elements is also associated with the occurrence of some neuromuscular diseases as well as cancer. We propose RNALocator, a neural network model to predict the cellular location of mRNAs. The features extracted from mRNA sequences along with the information gathered from their proteins are fed to this prediction model. We introduce a novel k-mer representation of sequences which is more memory and time efficient and can capture more information when k increases. The performance of RNALocator is compared to the results achieved by two powerful prediction models that were proposed in former studies named as mRNALoc and RNATracker on CeFra-seq and RNALocate datasets. The results reveal that the employment of protein-protein interaction, which plays a crucial role in many biological functions, together with the novel k-mer encoding of sequences, leads to a more accurate prediction model. However, the results also represent that by using only the k-mer information from mRNA sequences, RNALocator has a strong performance. On the other hand, RNALocator significantly reduces the computing time compared to previous studies. Source code and data used in this study are available at: `https://github.com/NeginBabaiha/RNALocator`.

Key words: Canonical k-mer representation, Distance-based k-mer, Protein-protein interaction

Introduction

The localization of RNA within the cell has an important role in the regulation of gene expression and has been suggested by numerous studies as a mechanism by which cells are polarized ([35]). The mRNA localization might be preferable to protein localization as one mRNA molecule can be like a template to code for plenty of proteins and thereby localizing mRNA rather than protein to the site where the protein is needed, is more efficient and time saving. The RNA localization is mainly carried out by a number of regulatory factors such as RNA-binding proteins that can control the distribution of mRNA in each compartment within the cells ([21, 24, 26, 47]). The relationship between the location of mRNAs and the proteins that they produce is usually complex and intertwined. Both types of molecules have signals in their sequence which can play a role in transferring them to specific locations within the cell ([6, 40]). Improper localization of RNA in the cell is also associated with the occurrence of some neuromuscular diseases and cancers ([7]). Formerly, a new class of drugs called oligonucleotides has been introduced that target the diseases-causing RNAs rather than proteins ([45, 20, 36]). Therefore, improving the understanding of the mRNA localization within the cell is beneficial in the fields such as drug design. Though mRNA localization has not been addressed as much as its importance, recently some computational predictors which are mainly based on machine learning approaches have been introduced. As an example, in the study by [54], a prediction model named as RNATracker is developed. RNATracker uses convolutional neural networks (CNN) combined with a special form of recurrent neural networks named as long short-term memory, briefly known as LSTM, to make prediction on CeFra-seq and APEX-RIP experimental data. CeFra-seq is a technology used to map human transcript abundance in

© The Author 2021. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

the nucleus, cytosol, endomembranes and insoluble fraction ([8]). APEX-RIP is an alternate technology that maps the transcriptome of the nucleus, cytoplasm, endoplasmic reticulum (ER) and mitochondria ([31]). Results achieved by RNATracker are better compared to the baseline neural network model named DNN-5Mer which is also studied and introduced by [54]. DNN-5Mer is based on popular k-mer representation and counts all 1-mers to 5-mers from the full RNA sequences and has two hidden layers. Moreover, RNATracker results gain no improvements when the secondary structure of mRNAs are also included in the model ([54]). Recently, in a study [53], a computational method named as RNA-GPS is introduced to predict localization of the transcripts that are achieved from APEX-RIP. For each transcript, RNA-GPS computes k-mer frequencies for k between 3 to 5 and assign a probability for each location using a random forest model. There is also another widely-used resource for RNA localization named as RNALocate ([55]). RNALocate is a web-accessible database that provides RNA sub-cellular localization resources and documents more than 190,000 RNA-associated sub-cellular localization entries with experimental and predicted evidence. It involves more than 105,000 RNAs with 44 sub-cellular locations in 65 species, mainly including Homo sapiens, Mus musculus, and Saccharomyces cerevisiae. The computational method mRNALoc which is recently investigated by [25], is mainly trained on a part of RNALocate dataset. mRNALoc extracts k-mer information from mRNA sequences and predicts cellular location of mRNAs using a support vector machine (SVM) ([25, 55]). As reported by [25], the performance of mRNALoc in two location of cytoplasm and nucleus is better comparing to a recently-introduced computational model named as iLoc-mRNA. iLoc-mRNA which is studied and designed by [56], is based on RNALocate dataset, too, and performs two strategies for feature selection and uses a SVM model for multi-class classification ([56]).

In this study, we introduce a novel k-mer based encoding for mRNA sequences which together with the canonical k-mer frequency matrix, can gather more sequence-based information from mRNAs. We develop a neural network based model, named as RNALocator that can also takes advantage of the protein-protein interactions (PPI) information. The PPI information is essential for many cellular functions. Proteins usually function through a number of interactions with other biomolecules and are widely used in biological problems to discover many functions ([50, 41, 37]). As an instance, in RNA metabolism and the context of discovering RNA binding proteins (RBPs), a successfull approach predicts RNA binding activity by analyzing protein-protein interactomes without relying on sequence or structure information ([9]). The importance of protein subcellular localization is due to the importance of protein's functions in different cell parts. Moreover, prediction of protein subcellular locations helps to identify the potential molecular targets for drugs and has an important role in genome annotation ([42]). All this information and the fact that the proteins with the similar protein-protein interaction patterns tend to be located mostly in the similar sub-cellular location([42], lead us to take advantage of including this valuable data in the predictive model.

Moreover, to develop a really useful predictor for a biological system, one can also follow "Chou's 5-steps rule" ([12]). Therefore, in this study, we go through the following five steps to establish the RNALocator predictor ([15, 17]):

1. Select or construct a valid benchmark dataset to train and test the predictor. This step is described in section 2.1.
2. Represent the mRNAs with an effective formulation and extract k-mer information that can reflect their intrinsic correlation with the target to be predicted. This step is described in 2.2.1 and 2.2.2.
3. Introduce and develop the powerful RNALocator algorithm to conduct the prediction. The three main steps of RNALocator and its work-flow are described in 2.3.
4. Properly perform cross-validation tests to objectively evaluate the anticipated prediction accuracy. This section is covered in the third step of RNALocator work-flow described in 2.3.
5. In our future work, we will establish a user-friendly web-server for the predictor that is accessible to the public. The more details are stated in the Conclusion section.

As mentioned in [12], the guidelines of Chou's 5-step rules have the following notable merits:

1. Crystal clear in logic development.
2. Completely transparent in operation.
3. Easily to repeat the reported results by other investigators.
4. With high potential in stimulating other sequence-analyzing methods.
5. Very convenient to be used by the majority of experimental scientist.

RNALocator is a neural network-based predictive model and also consistant with Chou's 5-steps. It is based on the Cefra-seq and RNALocate resources. The rest of the paper is organized as follows: In the Materials and Methods section, we first declare the datasets and also introduce the features that extracted from mRNA transcripts. Then, we describe the details and steps required to build our machine learning based model. In the Results section, we report the performance of RNALocator on the two mentioned datasets and also compare it with the following two state-of-art methods: mRNALoc and RNATracker. We also compare the performance of 10-fold cross-validation by RNALocator with a baseline method named as DNN-5Mer which is formerly executed on Cefra-seq dataset and hence assess the importance of the features applied in RNALocator. Finally, in the Discussion section, we extend our experiments to non-human related transcripts and also assess the performance of RNALocator using both novel and canonical k-mer information on both human and non-human transcripts.

Materials and Methods

Here, we introduce the data sources employed in our research. Then, the details of the information and features are explained and finally the architecture and methodological aspects of RNALocator are described in depth.

Data Sources

mRNA Sequences and Localization Information

Two benchmark datasets are considered. The first dataset is CeFra-seq which is also employed in RNATracker method. As mentioned above, Cefra-seq contains the human transcripts and localization information of mRNAs is presented in the form of normalized gene expression values in each of four sub-cellular locations: cytosol, nuclear, membranes, and insoluble. Hence, for each mRNA, rather than a single cellular location label, we have a vector of length 4 whose elements represent the probability of belonging to each location. The total number of mRNAs is 11,373 and the sequences are downloaded from the Ensemble database ([1]).

In the second dataset, mRNA sequences and sub-cellular localization information are extracted from RNALocate dataset. The same training dataset employed in mRNALoc method is used

Table 1. Total number of mRNAs in each five locations in RNALocate dataset. The first column represents each cellular compartment. The second and third column reveal the number of human and non-human transcripts, respectively.

Location	Human Species	Non-human species
Cytoplasm	3,419	1,509
Reticulum Endoplastic	1,168	6
Extracellular Reigion	26	506
Mitochondria	2	344
Nuclear	1,265	2,869
Total	5,880	5,244

Table 2. Selected hyperparameters for RNALocator.

Parameter	Value
Epochs	300
Batch size	512
Hidden layer	1
Fully connected neurons	200
Dropout rate	0.2

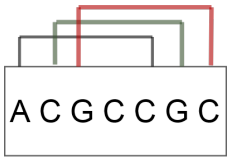

Fig. 1. All 5-mer structures contained in ACGCCGC sequence. In this example, we have three 5-mers ACGCC, CGCCG and GCCGC that are shown in three different colors.

in this paper, however, RNALocator considers each human and non-human transcripts separately and also for each gene only one isoform is taken into account. In this study only the mRNA sequences that belong to a single location are considered, and it covers five sub-cellular compartments: cytoplasm, endoplasmic reticulum, extracellular region, mitochondrial, and nucleus. The total number of mRNAs here is 11,124 where 5,880 of them are human transcripts and 5,244 of them belong to non-human transcripts. The summary of this dataset is shown in Table 1.

Protein-Protein Interaction (PPI) Information

The PPI information for human-related mRNAs is extracted from STRING database ([51]). In this database, it usually considers the longest protein-coding isoform among all isoforms of a gene. Hence, one protein is assigned to each mRNA and we keep this information in the form of an adjacency matrix. Each element of this matrix represents the strength of the interaction between two proteins which is assigned by STRING.

Feature encoding

Two types of features are adopted from the mRNA sequences. The first one is k-mer representation which is one of the most frequently used encodings for nucleotide sequences ([3, 28, 33]). The second one is a novel suggested k-mer representation which is based on the distance. The details of these features are described as follows:

(5-mer) representation

With the explosive growth of biological sequences in the post-genomic era, one of the most important but also most difficult problems in computational biology is how to express a biological sequence with a discrete model or a vector, yet still keep considerable

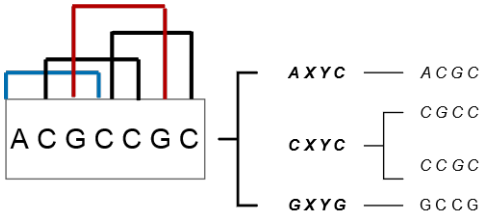

Fig. 2. All distance-based 2-mer structures contained in ACGCCGC sequence. In this example, we have 2 substructure of form: $CxyC$, 1 of form $AxyC$ and also 1 of form $GxyG$. Each three distance-based 2-mers are shown in three different colors.

sequence-order information or key pattern characteristics. This is because all the existing machine-learning algorithms can only handle vectors as elaborated in a comprehensive review ([13]). One of the ways to extract a uniform length feature vector from these sequences is to count k-mer frequencies ([34, 3]). A k-mer refers to a possible subsequences of length k in a mRNA sequence. As we have 4 neucliotides, the total number of possible k-mers is 4^k.

As an example, for the ACGCCGC sequence, all 5-mer structures are: ACGCC, CGCCG, and GCCGC that are also demonstrated in Figure 1. This k-mer features extracted in this paper can be covered by a very powerful web-server called "Pse-in-One([39]).

For a mRNA sequence S, we first sort all the 5-mer structures lexicographically, then count the occurrence of each one in the main mRNA sequence and get the following feature vetor: $F_5(S) = [v_1, v_2, ..., v_n]$ where each v_i is the frequency of i-th 5-mer, and in this case n is equal to $4^5 = 1024$.

Distance-based k-mer frequencies

One of the drawbacks of k-mer representation is that when k increases, the feature vector becomes extremely large and sparse which can be memory-inefficient and can reduce the performance of model. In this work, to overcome this problem and to extract more information with a smaller feature vector for large values of k, we propose a novel distance-based k-mer representation. In this representation, the distance between the first and last nucleotide of the subsequence that we count, is set to be k. Then, for each two nucleotides coming with distance k from each other, the frequency of this subsequence is computed. So, for a mRNA sequence S and the distance k, we get the following feature vector of length 16: $D_k(S) = [w_1, w_2, ..., w_{16}]$, where each w_i is the frequency of the each distance-based k-mer. As an example, for k = 2, all subsequences to count are illustrated in bellow:

$$
\begin{array}{llll}
w_1 : AxyA & w_2 : AxyC & w_3 : AxyG & w_4 : AxyT \\
w_5 : CxyA & w_6 : CxyC & w_7 : CxyG & w_8 : CxyT \\
w_9 : GxyA & w_{10} : GxyC & w_{11} : GxyG & w_{12} : GxyT \\
w_{13} : TxyA & w_{14} : TxyC & w_{15} : TxyG & w_{16} : TxyT
\end{array}
$$

Here, x and y can be substituted with any of four following nucleotides: A, G, C, and T. As an example, let's consider the S to be the mRNA with sequence ACGCCGC. In this sequence, for $w_2 : AxyC$, only ACGC is contained in the sequence, so frequency of w_2 is 1. For w_6 which is $CxyC$, two subsequences of $CGCC$ and $CCGC$ are contained in the mRNA sequence and hence the frequency of w_6 is 2. For w_{11} which is $GxyG$, only the subsequence $GCCG$ is observed and therefore the frequency of w_{11} is 1. Other w_is have frequency of zero in this sequence. This example is also illustrated in Figure 2. After experimenting with a variety of ranges

for k, we set k to be in range 0 to 8 hence coming up with a feature vector of length $9 * 16 = 144$.

Principle Component Analysis on PPI network

As mentioned above, the PPI information is reflected in the form of the adjacency matrix with the dimention of: number of mRNAs $*$ number of mRNAs. Therefore, for the first dataset, the PPI matrix has a dimension of $11,373 * 11,373$, and for the human mRNAs in RNALocate dataset, it is a matrix of size of $5880 * 5880$. As performance of machine learning models can degrade when too many features are considered, we first reduce the dimension of this matrix using Principle Component Analysis (PCA) technique ([30]). PCA is among the most widely used methods to reduce the feature space and improve storage space or the computational efficiency of the learning algorithm. It uses singular value decomposition to project the data it to a lower dimensional space and emphasizes variation and brings out strong patterns in a dataset. In this study, after some experiments, the number of principle components is set to 500 and the total explained variance is more than 70 percent of the whole data this way.

RNALocator

In this section, we express the main steps of RNALocator.
Step 1: Aggregation of the following three feature vectors:

1. 5-mer frequencies (a vector of length 1024)
2. Distance-based k-mer frequencies (a vector of length 144)
3. Reduced PPI matrix using PCA method (a vector of length 500 for each mRNA)

We concatenate the gathered information into a single feature vector of dimension $(1024 + 144 + 500)$. This vector is the final vector which is further fed into our Neural Network model for the prediction task.

Step 2: Design of a neural network model

Based on the features obtained, we propose an artificial neural network (ANN) to assign a probability to each of the location classes for each mRNA. A neural network can be expressed as a series of matrix multiplications interleaved with non-linear function. ANN consists of a number of smaller units called neurons which can be repeated in many layers. To prevent the neural network from getting sophisticated and thus becoming more difficult to be trained efficiently, we design a shallow architecture for our model which has one hidden layer and 200 neurons. In the hidden layer, dropout is also used to randomly mask half of the connections during the training of the model to prevent overfitting. We use the rectified linear unit (Relu) activation function in the hidden layer which is defined as below ([23]):

$$\text{Relu(x)} = max(0, x). \tag{1}$$

The Softmax function as the non-linear function is applied in the last layer of the model to assign a probability to each location and is formulated as bellow ([49]):

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{2}$$

Finaly, we use Kullback-Leiber-Divergence as the loss function. For probability distribution $\mathfrak{P}$ and $\mathfrak{Q}$ defined on the same probability space X, Kullback-Leiber-Divergence is defined as ([52]):

$$KL(\mathfrak{Q}\|\mathfrak{P}) = \Sigma_{x \in X}(log\frac{\mathfrak{Q}(x)}{\mathfrak{P}(x)}). \tag{3}$$

Step 3: Training of the prediction model

Hyper-parameters are chosen empirically based on the training dataset. We chose all parameters with the purpose to minimize the loss function. Briefly, we employ the 10-fold cross-validation mode for training the model. Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it ([5]. More specifically, In k-fold cross-validation, the data is first partitioned into k equally (or nearly equally) sized segments or folds. Consequently, k iterations of training and validation are performed such that within each iteration a different fold of the data is held-out for validation while the remaining $k - 1$ folds are used for learning ([5]). We then evaluate the results using a range of values for hidden layers (No hidden layer, 1, 2 and 3), for neurons in each fully connected layer (1000, 700, 500, 200, 100) and for dropout rates (0.1, 0.2, 0.3 and 0.5). The best parameters that are used in this model are shown in Table 2. The number of epochs for training each fold is 300 and a validation set consisting of 10 percent of the training data is also applied to monitor the loss function in the training process and detect overfitting. RNALocator is implemented using the Keras Library ([29]). Adam optimizer with Nesterov momentum is also employed for training the model ([22]). The overall workflow of RNALocator is shown in Figure 3.

Results

Time complexity

Many of the methodologies like mRNALoc, are available as a server tool and hence make it impossible to do time complexity related comparisons. Among them, however, RNAtracker's source code and implementation are available and RNALocator outperforms considerably the RNAtracker method in terms of training time. For RNALocator, this time is computed as about 3 hours on Linux Ubuntu machine, with 15 CPUs(Intel Xeon(R) 2.00 GHz) on CeFra-seq dataset and comparing to RNATracker, is noticeably reduced(RNAtracker requires 7 days for training in full length mode and 8 hours in fixed length mode on GTX1080Ti graphic card). On the RNALocate dataset, RNALocator has training time of 2 hours which is again remarkably decreased comparing to RNATracker which needs the computation time of 6 hours. The computation time of DNN-5Mer is roughly the same as RNALocator.

Evaluation criteria

A pearson correlation is computed to compare the predicted values with experimental ones. Pearson correlation is a measure to assess the linear correlation between variables. It has a value in range $(-1, +1)$. A value of $+1$ is total positive linear correlation, 0 is no linear correlation, and -1 is total negative linear correlation. To better evaluate the model's performance, we also consider the Spearman correlation between the predicted and experimental values to capture the order of locations to which a mRNA belongs. For the second dataset, as one location is dedicated to each mRNA, we report the location with the maximum probability to be the final predicted location of mRNA. In order to compare the performance of RNALocator method to other methods for second dataset, True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Precision, Recall, F-score, Accuracy (ACC) and

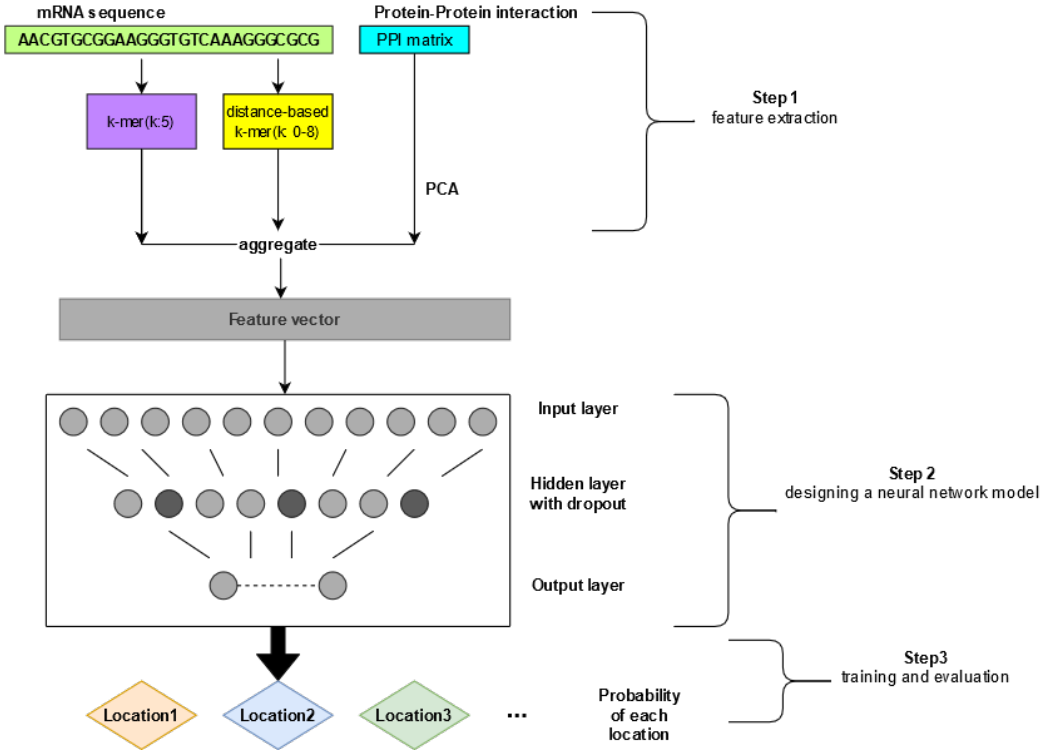

Fig. 3. The overflow of RNALocator method. In Step 1, we first aggregate all information gathered from both sequence-based features as well as PPI matrix. In Step 2, we design a neural network model with the mentioned architecture. In Step 3, the model is trained and evaluated using 10-fold cross-validation. For the CeFra-seq dataset, we report a vector of length 4 for each mRNA and for the RNALocate dataset, we report the location with maximum probability to be the predicted location of the mRNA.

Matthews Correlation Coefficient (MCC) are calculated as follows:

$$
\begin{aligned}
Precision &= \frac{TP}{TP+FP},\ Recall = \frac{TP}{TP+FN},\\
F-score &= 2\frac{Recall * Precision}{Recall + Precision},\\
ACC &= \frac{TP+TN}{TP+FP+FN+TN},\\
MCC &= \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}.
\end{aligned}
$$

The criteria mentioned above are among the most commonly used metrics in classification problems that can define how good a model has performed.

Assessment and Comparison

The average results of 30 times 10-fold cross-validation mode of training on the first dataset are shown in Table 3. In the CeFra-seq dataset, RNALocator obtains the Pearson correlations of 0.69, 0.65, 0.54 and 0.55 in cytosol, insoluble, membrane and nuclear, respectively. These correlations are higher in all locations comparing to RNATracker fixed length mode method. For total 2893 number of mRNAs in this dataset, the Spearman correlation between the real and predicted vectors is computed as 1 which is again slightly better comparing to RNATracker(2849). As Table 3 demonstrates, RNALocator achieves about 16% better overall Pearson correlation comparing to RNATracker fixed length mode and about 1% better overall Pearson correlation compared to RNATracker full-length mode. As mRNALoc is available as a standalone tool which is trained based on five different locations than Cefra-seq, it was not feasible to take it into the comparison with RNALocator on this dataset. We also compare the performance of RNALocator with DNN-5Mer. As mentioned in the former sections, DNN-5Mer extracts k-mer features and has only two hidden layers with the same number of neurons as the input vector. It also uses Relu activation function in the hidden layer. Though both RNALocator and DNN-5Mer have a simple architecture, however, the performance of DNN-5Mer is noticeably weak and it achieves the Pearson correlation of 0.63 in cytosol, 0.55 in insoluble, 0.42 in membrane and 0.48 in nuclear and in overal, RNALocator achieves about 34 percent higher Pearson correlation comparing to DNN-5Mer. The results can reveal the benefits of employing the most relative and informative features in the prediction model when designing a shallow neural network architecture.

The performance of RNALocator on the second dataset is also shown in the Table 4. As illustrated in this table, in cytosol RNALocator has Precision, Recall and F-score of 74%, 72%, and 74%, respectively. In Endoplasmic reticulum , the Precision is 56% and Recall is 48% while F-score is 52% .In Extracellular region and mitochondria, due to few training samples(only 26 and 2, respectively) belonging to these location, Recall and F-score are all near to zero and the Precision in Extracellular region is 91%. In nucleus, the Precision of prediction is 52%, while Recall and F-score are 70% and 60%, respectively. In three following locations cytosol, ER and EX, RNALocator has improved the average of F-score in all locations by 4% comparing to mRNALoc, and by 12% comparing to RNATracker. In nucleus, however, the average of F-score obtained for RNALocator is pretty much the same as mRNALoc. The overall Accuracy of prediction using RNALocator is 65%, which is 4% better comparing to mRNALoc and 2% higher when compared to

Table 3. Results of 30 10-fold cross-validation of three methods: RNALocator(with and without employing PPI), RNATracker(fixed and full length mode) and DNN-5Mer on CefraSeq daatset.

Location	RNALocator	RNALocator(no PPI)	RNATracker fixed	RNATracker full	DNN-5Mer
Cytosol	0.69	0.67	0.68	**0.70**	0.63
Insoluble	**0.65**	0.61	0.62	0.64	0.55
Mebrane	**0.54**	0.52	0.47	**0.54**	0.42
Nuclear	**0.55**	0.52	0.50	0.54	0.48

RNATracker(fixed length mode). RNALocator achieves also MCC of 0.40, which is higher comparing to both RNATracker and mRNALoc methods(that achieve MCC of 0.34 and 0.37, respectively.)

Discussion

In order to assess the impact of including the PPI information in our model, we perform the following analysis on both Cefra-Seq and RNALocate datasets. Here, we employ both 5-mer and distance-based k-mer information gathered from mRNA sequences and then, compare the results with the case when PPI information is also used in the model. As Table 3 demonstrates, on Cefra-se dataset when RNALocator employs the reduced PPI matrix in the model, it performs significantly better and achieves 10% higher Pearson correlation in average for all locations. However, it is also important to notice that by using the 5-mer and distance-based k-mer information and without inclusion of PPI, in cytosol and insoluble it achieves almost the same Pearson correlation compared to RNATracker(fixed length), and in membrane and nucleus it achieves 5 and 2 percent higher correlation, individually. This makes it clear that even without taking advantage the PPI information in the model, RNALocator outperforms the former method RNATracker and achieves higher overall Pearson correlation. We do the same analysis on the human-related transcripts of RNALocate dataset and consider both 5-mer and distance-based the k-mer information in the model. As Table 4 demonstrates, in this case the average of Precision, Recall, F-score, Accuracy and MCC in all locations are computed as 0.33, 0.29 and 0.30, 0.63 and 0.30, respectively. These results represent that in the second dataset, too, when RNAlocator employs PPI information, the performance gets clearly better and it achieves average of Precision, Recall, F-score, Accuracy and MCC of 0.55, 0.38, 0.37, 0.65 and 0.40, respectively.

This improvement in the performance of RNALocator when it considers the PPI information, may probably highlight the importance of this PPI information and reveal its potential role cellular localization of mRNAs. However, it is also noticeable that on human transcripts of RNALocate dataset, even without including PPI information, RNALocator achieves higher average Accuracy of prediction comparing to both mRNALoc and RNATracker methods. RNALocator also has the average F-score of 0.30 and average Recall of 0.30 in all locations in this case which is again better comparing to RNATracker method that has average F-score and Recall of 0.24 and 0.28, respectively.

At last, we also consider the non-human transcripts of RNALocate dataset. In this case too, we employ both 5-mer and distance-based k-mer information of mRNA sequences in the model. The performance of RNALocator on non-human species is described in Table 5 and is compared to mRNALoc and RNATracker. The details and distributions of non-human transcripts of RNALocate dataset are formerly declared in Table 1. The average of Precision, Recall, F-score, ACC and MCC in all locations for RNALocator in this case are 3%, 6%, 6%, 4% and 8% higher comparing to RNATracker, respectively. Comparing to mRNALoc method, too, the average of Precision, ACC and MCC for RNALocator are 6%, 9% and 12% better, separately.

Eventually, as illustrated by the results, RNALocator has been capable to outperform both methods in most of the criteria in these two datasets. This is determinant of the strong power of RNALocator while it has a simple and transparent neural network architecture. Besides, it's noticeable to bare in mind that the pre-trained mRNALoc software with which we compared our results on the second dataset, is already trained on this dataset and nonetheless the better perfomance of RNALocator is highly meritorious.

Conclusion

RNALocator stands among the few proposed methods that have applied the machine learning based approaches to study the cellular localization of mRNAs. Facing the explosive growth of biological sequences discovered in the post-genomic age, to timely use them for drug development, a lot of important sequence-based information, such as PTM (posttranslational modification) sites in proteins, have been successfully predicted ([18]). Actually, the rapid development in sequential bioinformatics and structural bioinformatics have driven the medicinal chemistry undergoing an unprecedented revolution ([14]), in which the computational biology has played an increasingly important role in stimulating the development of finding novel drugs. In view of this, the computational (or in silico) methods were also utilized in the current study. RNALocator aims to profit from the interaction information of the proteins coded by mRNA transcripts, as this information has long been studied and employed in a variety of problems such as analyzing genome reductive evolution or identifying disease causing genes ([2, 43]). First and foremost, we demonstrated that a simple neural network model can be effective as well when the most informative features are brought into the model. This simplicity also reduces the required computational time of the training models to a remarkable extent. It is also important to notice that the Accuracy of prediction is improved when a dimensional reduction technique such as PCA is applied on the PPI matrix. In the future, other dimensionality reduction techniques can be employed such as autoencoder compressing approach. We declared that the distance-based k-mer representation was introduced to address the limitation in canonical k-mer frequency computation that when parameter k is increased, the dimension of the obtained feature vector becomes too large and the probability of observing any specific k-mer becomes very small. Besides, as the frequently used k-mer representations can cause memory-related problems when k increases, we benefited from a novel k-mer representation based on the distance which was less prone to add redundancy to the data when k is increased. Moreover, it's precious to notice that inclusion of more relative information could also be further assessed in the future studies. As an example, the knowledge of protein 3D (three-dimensional) structures or their complexes with ligands which is vitally important for rational drug design ([27]). Although X-ray crystallography is a powerful tool in determining these structures ([32]), it is time-consuming and expensive, and not all proteins can be successfully crystallized. Membrane proteins are difficult to crystallize and most of them will not dissolve in normal solvents. Therefore, so far very few membrane protein structures have been determined. NMR

Table 4. Results of 30 10-fold cross-validation of RNALocator (with and without employing PPI) compared with RNATracker(fixed length mode) and mRNALoc on the human part of RNALocate dataset. The names of compartments are abbreviated as Cyt : cytosol, ER: Endoplasmic reticulum, EX : Extracellular region, Mit :Mitochondria, Nuc: Nucleus.

	RNALocator					RNALocator(noPPI)					RNATracker					mRNALoc				
Criteria	Cyt	ER	EX	Mit	Nuc	Cyt	ER	EX	Mit	Nuc	Cyt	ER	EX	Mit	Nuc	Cyt	ER	EX	Mit	Nuc
Precision	**0.74**	**0.56**	**0.91**	0.00	0.52	0.65	0.40	0.00	0.00	**0.61**	0.66	0.52	0.00	0.00	0.49	0.74	0.46	0.00	0.00	0.52
Recall	0.72	**0.48**	**0.04**	0.00	0.70	**0.84**	0.14	0.00	0.00	0.51	0.59	0.06	0.00	0.00	0.54	0.61	0.50	0.00	0.00	**0.72**
F-score	**0.74**	**0.52**	**0.07**	0.00	0.60	0.73	0.21	0.00	0.00	0.56	0.73	0.11	0.00	0.00	0.53	0.67	0.48	0.00	0.00	**0.61**
Average Accuracy	**0.65**					0.63					0.63					0.61				
Average MCC	**0.40**					0.30					0.34					0.37				

Table 5. Performance of RNALocator, RNATracker(fixed length mode) and mRNALoc on non-human mRNAs of RNALocate dataset.

	RNALocator					mRNALoc					RNATracker				
	Cyt	ER	EX	Mit	Nuc	Cyt	ER	EX	Mit	Nuc	Cyt	ER	EX	Mit	Nuc
Precision	**0.77**	0.00	**0.38**	0.98	0.72	0.57	**0.01**	0.23	**0.99**	**0.77**	0.69	0.00	0.50	0.85	0.70
Recall	0.54	0.00	0.03	**0.98**	**0.96**	**0.57**	**0.33**	**0.18**	0.85	0.76	0.54	0.00	0.001	0.75	0.92
F-score	**0.63**	0.00	0.05	**0.98**	**0.82**	0.57	**0.02**	**0.20**	0.92	0.77	0.60	0.00	0.003	0.79	0.79
Average Accuracy	**0.74**					0.65					0.70				
Average MCC	**0.55**					0.43					0.47				

is indeed a very powerful tool in determining the 3D structures of membrane proteins([48, 4, 44, 46, 19]), but it is also time-consuming and costly. To acquire the structural information in a timely manner, a series of 3D protein structures have been developed by means of structural bioinformatics tools ([11]).Therefore, in the future versions of RNALocator, this information could be taken into account which could hopefully improve the performance of RNALocator. Besides, as shown in a series of recent publications ([10, 38]) in demonstrating new findings or approaches, user-friendly and publicly accessible web-servers will significantly enhance their impacts ([16, 12]), and so we shall make efforts in our future work to provide a web-server to display the findings that can be manipulated by users according to their needs.

Acknowledgments

Changiz Eslahchi and others would like to thank the School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM) and Computing Center of IPM in performing a parallel computing is gratefully acknowledged.

References

1. Bronwen L Aken, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Friederike Bernsdorff, Jyothish Bhai, Konstantinos Billis, Denise Carvalho-Silva, Carla Cummins, Peter Clapham, et al. Ensembl 2017. *Nucleic acids research*, 45(D1):D635–D642, 2017.
2. Richard O Akinola, Gaston K Mazandu, and Nicola J Mulder. A quantitative approach to analyzing genome reductive evolution using protein–protein interaction networks: a case study of mycobacterium leprae. *Frontiers in genetics*, 7:39, 2016.
3. Ehsaneddin Asgari, Kiavash Garakani, Alice C McHardy, and Mohammad RK Mofrad. Micropheno: predicting environments and host phenotypes from 16s rrna gene sequencing using a k-mer based representation of shallow sub-samples. *Bioinformatics*, 34(13):i32–i42, 2018.
4. Marcelo J Berardi, William M Shih, Stephen C Harrison, and James J Chou. Mitochondrial uncoupling protein 2 structure determined by nmr molecular fragment searching. *Nature*, 476(7358):109–113, 2011.
5. Daniel Berrar. Cross-validation., 2019.
6. Günter Blobel. Intracellular protein topogenesis. *Proceedings of the National Academy of Sciences*, 77(3):1496–1500, 1980.
7. Louis Philip Benoit Bouvrette, Neal AL Cody, Julie Bergalet, Fabio Alexis Lefebvre, Cédric Diot, Xiaofeng Wang, Mathieu Blanchette, and Eric Lécuyer. Cefra-seq reveals broad asymmetric mrna and noncoding rna distribution profiles in drosophila and human cells. *Rna*, 24(1):98–113, 2018.
8. Louis Philip Benoit Bouvrette, Neal AL Cody, Julie Bergalet, Fabio Alexis Lefebvre, Cédric Diot, Xiaofeng Wang, Mathieu Blanchette, and Eric Lécuyer. Cefra-seq reveals broad asymmetric mrna and noncoding rna distribution profiles in drosophila and human cells. *Rna*, 24(1):98–113, 2018.
9. Kristopher W Brannan, Wenhao Jin, Stephanie C Huelga, Charles AS Banks, Joshua M Gilmore, Laurence Florens, Michael P Washburn, Eric L Van Nostrand, Gabriel A Pratt, Marie K Schwinn, et al. Sonar discovers rna-binding proteins from analysis of large-scale protein-protein interactomes. *Molecular cell*, 64(2):282–293, 2016.
10. Wei Chen, Hua Tang, Jing Ye, Hao Lin, and Kuo-Chen Chou. irna-pseu: Identifying rna pseudouridine sites. *Molecular Therapy-Nucleic Acids*, 5:e332, 2016.
11. Kuo-Chen Chou. Structural bioinformatics and its impact to biomedical science. *Current medicinal chemistry*, 11(16):2105–2134, 2004.
12. Kuo-Chen Chou. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of theoretical biology*, 273(1):236–247, 2011.
13. Kuo-Chen Chou. Impacts of bioinformatics to medicinal chemistry. *Medicinal chemistry*, 11(3):218–234, 2015.
14. Kuo-Chen Chou. Impacts of bioinformatics to medicinal chemistry. *Medicinal chemistry*, 11(3):218–234, 2015.
15. Kuo-Chen Chou. Advances in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. *Current medicinal chemistry*, 26(26):4918–4943, 2019.
16. Kuo-Chen Chou. Advances in predicting subcellular localization of multi-label proteins and its implication for developing multi-target drugs. *Current medicinal chemistry*, 26(26):4918–4943, 2019.

17. Kuo-Chen Chou. Impacts of pseudo amino acid components and 5-steps rule to proteomics and proteome analysis. *Current topics in medicinal chemistry*, 19(25):2283–2300, 2019.
18. Kuo-Chen Chou. Progresses in predicting post-translational modification. *International Journal of Peptide Research and Therapeutics*, pages 1–16, 2019.
19. Jyoti Dev, Donghyun Park, Qingshan Fu, Jia Chen, Heather Jiwon Ha, Fadi Ghantous, Tobias Herrmann, Weiting Chang, Zhijun Liu, Gary Frey, et al. Structural basis for membrane anchoring of hiv-1 envelope spike. *Science*, 353(6295):172–175, 2016.
20. Scott J Dixon and Brent R Stockwell. Identifying druggable disease-modifying gene products. *Current opinion in chemical biology*, 13(5-6):549–555, 2009.
21. Daniel Dominguez, Peter Freese, Maria S Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, Nicole J Lambert, Eric L Van Nostrand, Gabriel A Pratt, et al. Sequence, structure, and context preferences of human rna binding proteins. *Molecular cell*, 70(5):854–867, 2018.
22. Timothy Dozat. Incorporating nesterov momentum into adam. 2016.
23. Konstantin Eckle and Johannes Schmidt-Hieber. A comparison of deep networks with relu activation function and linear spline-type methods. *Neural Networks*, 110:232–242, 2019.
24. Fabrizio Ferre, Alessio Colantoni, and Manuela Helmer-Citterich. Revealing protein–lncrna interaction. *Briefings in bioinformatics*, 17(1):106–116, 2016.
25. Anjali Garg, Neelja Singhal, Ravindra Kumar, and Manish Kumar. mrnaloc: a novel machine-learning based in-silico tool to predict mrna subcellular localization. *Nucleic Acids Research*, 2020.
26. Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. A census of human rna-binding proteins. *Nature Reviews Genetics*, 15(12):829–845, 2014.
27. Jonathan Greer, John W Erickson, John J Baldwin, and Michael D Varney. Application of the three-dimensional structures of protein target molecules in structure-based drug design. *Journal of medicinal chemistry*, 37(8):1035–1054, 1994.
28. Brian L Gudenas and Liangjiang Wang. Prediction of lncrna subcellular localization with deep learning from sequence features. *Scientific reports*, 8(1):1–10, 2018.
29. Antonio Gulli and Sujit Pal. *Deep learning with Keras*. Packt Publishing Ltd, 2017.
30. Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.
31. Pornchai Kaewsapsak, David Michael Shechner, William Mallard, John L Rinn, and Alice Y Ting. Live-cell mapping of organelle-associated rnas via proximity biotinylation combined with protein-rna crosslinking. *Elife*, 6:e29224, 2017.
32. Ali A Kermani. A guide to membrane protein x-ray crystallography. *The FEBS journal*, 2020.
33. Jessime M Kirk, Susan O Kim, Kaoru Inoue, Matthew J Smola, David M Lee, Megan D Schertzer, Joshua S Wooten, Allison R Baker, Daniel Sprague, David W Collins, et al. Functional classification of long non-coding rnas by k-mer content. *Nature genetics*, 50(10):1474–1482, 2018.
34. Jessime M Kirk, Susan O Kim, Kaoru Inoue, Matthew J Smola, David M Lee, Megan D Schertzer, Joshua S Wooten, Allison R Baker, Daniel Sprague, David W Collins, et al. Functional classification of long non-coding rnas by k-mer content. *Nature genetics*, 50(10):1474–1482, 2018.
35. Malgorzata Kloc, N Ruth Zearfoss, and Laurence D Etkin. Mechanisms of subcellular mrna localization. *Cell*, 108(4):533–544, 2002.
36. Arthur A Levin. Treating disease at the rna level with oligonucleotides. *New England Journal of Medicine*, 380(1):57–70, 2019.
37. Dajeong Lim, Nam-Kuk Kim, Hye-Sun Park, Seung-Hwan Lee, Yong-Min Cho, Sung Jong Oh, Tae-Hun Kim, and Heebal Kim. Identification of candidate genes related to bovine marbling using protein-protein interaction networks. *International journal of biological sciences*, 7(7):992, 2011.
38. Bin Liu, Longyun Fang, Ren Long, Xun Lan, and Kuo-Chen Chou. ienhancer-2l: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics*, 32(3):362–369, 2016.
39. Bin Liu, Fule Liu, Xiaolong Wang, Junjie Chen, Longyun Fang, and Kuo-Chen Chou. Pse-in-one: a web server for generating various modes of pseudo components of dna, rna, and protein sequences. *Nucleic acids research*, 43(W1):W65–W71, 2015.
40. Kelsey C Martin and Anne Ephrussi. mrna localization: gene expression in the spatial dimension. *Cell*, 136(4):719–730, 2009.
41. Suresh Mathivanan, Balamurugan Periaswamy, TKB Gandhi, Kumaran Kandasamy, Shubha Suresh, Riaz Mohmood, YL Ramachandra, and Akhilesh Pandey. An evaluation of human protein-protein interaction data in the public domain. In *BMC bioinformatics*, volume 7, page S19. Springer, 2006.
42. Elnaz Mirzaei Mehrabad, Reza Hassanzadeh, and Changiz Eslahchi. Pmlpr: A novel method for predicting subcellular localization based on recommender systems. *Scientific reports*, 8(1):1–10, 2018.
43. Martin Oti, Berend Snel, Martijn A Huynen, and Han G Brunner. Predicting disease genes using protein–protein interactions. *Journal of medical genetics*, 43(8):691–698, 2006.
44. Bo OuYang, Shiqi Xie, Marcelo J Berardi, Xinhao Zhao, Jyoti Dev, Wenjing Yu, Bing Sun, and James J Chou. Unusual architecture of the p7 channel from hepatitis c virus. *Nature*, 498(7455):521–525, 2013.
45. John P Overington, Bissan Al-Lazikani, and Andrew L Hopkins. How many drug targets are there? *Nature reviews Drug discovery*, 5(12):993–996, 2006.
46. Kirill Oxenoid, Ying Dong, Chan Cao, Tanxing Cui, Yasemin Sancak, Andrew L Markhard, Zenon Grabarek, Liangliang Kong, Zhijun Liu, Bo Ouyang, et al. Architecture of the mitochondrial calcium uniporter. *Nature*, 533(7602):269–273, 2016.
47. Debashish Ray, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, et al. A compendium of rna-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, 2013.
48. Jason R Schnell and James J Chou. Structure and mechanism of the m2 proton channel of influenza a virus. *Nature*, 451(7178):591–595, 2008.
49. Sagar Sharma. Activation functions in neural networks. *Towards Data Science*, 6, 2017.
50. Daniela Stojanova, Michelangelo Ceci, Donato Malerba, and Saso Dzeroski. Using ppi network autocorrelation in hierarchical multi-label classification trees for gene function prediction. *BMC bioinformatics*, 14(1):285, 2013.
51. Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page

gkw937, 2016.

52. Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.
53. Kevin E Wu, Kevin R Parker, Furqan M Fazal, Howard Chang, and James Zou. Rna-gps predicts high-resolution rna subcellular localization and highlights the role of splicing. *RNA*, pages rna–074161, 2020.
54. Zichao Yan, Eric Lécuyer, and Mathieu Blanchette. Prediction of mrna subcellular localization using deep recurrent neural networks. *Bioinformatics*, 35(14):i333–i342, 2019.
55. Ting Zhang, Puwen Tan, Liqiang Wang, Nana Jin, Yana Li, Lin Zhang, Huan Yang, Zhenyu Hu, Lining Zhang, Chunyu Hu, et al. Rnalocate: a resource for rna subcellular localizations. *Nucleic acids research*, 45(D1):D135–D138, 2017.
56. Zhao-Yue Zhang, Yu-He Yang, Hui Ding, Dong Wang, Wei Chen, and Hao Lin. Design powerful predictor for mrna subcellular location prediction in homo sapiens. *Briefings in Bioinformatics*, 2020.